WATER CHESS - Group 20
HI!CKATHON 2023

14/01/2023

OVERVIEW

1. Project Background and Description

As a group of conscious and progress-oriented students, we decided to assemble our Data and Business skills and hop on an impactful venture. Our project aims at encouraging self-renovation in French single houses, through accurate energy usage predictions thanks to our Machine Learning model and a user-friendly interactive app. We are excited to see the positive impact we can create, and hope investors and experts will join us in this adventure.

2. Project Scope

The final product of this project is a machine learning model which will estimate the energy consumption of the European buildings based on their various features and information about yearly energy consumption. The large dataset will be analyzed, and appropriate data preprocessing will be done to prepare the input for the model training. We will start with simple preprocessing techniques and apply more complicated ones during the optimization step depending on the time constraints. Once the final model is created we will show how it can be leveraged for energy and cost savings purposes. This goal will be achieved by presenting our market analysis and economic/societal impact in a short video business pitch.

3. Presentation of the group

Include your specialization at school, etc.

First name	Last name	Year of studies & profile	School	Skills	Roles/Tasks	Observations
Stefano	Bavaro	M1 & AI	UPSaclay	AI/ML, data analysis	data preprocessing, model optimization	
Matteo Roberto	Facta	M1 & Management	HEC	Economics, management, statistics, Finance	Business planning - Business Pitch	
Khuong	Thanh Gia Hieu	M1 AI	UPSaclay	AI/ML, data analysis	data preprocessing, model creation and optimization	

Benedetta	Magni	M1 & Management	HEC	Management, statistics, design, Finance	Business planning - Business Pitch - Video making and editing	
Diego Andres	Torres Guarin	M1 A1	UPSaclay	AI/ML, data analysis	data preprocessing, model optimization	
Dana	Zhumabekova	M2 & ROSP	IP Paris	AI/ML, data analysis	data preprocessing, model optimization	
Mahdi	Ranjbar	M1 & AI	UPSaclay	AI/ML, data analysis	data preprocessing, model optimization	

4. **Task Management**

We discussed first the purpose of our project, business goals and the Data Science approach we will use to achieve them. We looked together at the dataset, and discussed the data preprocessing part, which was one of the essential steps. Then the work was split according to our skills. Stefano, Khuong, Diego and Dana worked on data preprocessing by dividing the features. Khuong developed the ML model, and then we all contributed to the model optimization and further data cleaning. Matteo and Benedetta worked on the business part of the project: developing the business idea, market analysis, designing the app and creating a video. Afterwards they presented the business model to everyone, and we all discussed and gave our opinion. It should be mentioned that throughout the whole project we constantly collaborated and shared the ideas.

PROJECT MANAGEMENT

1. **Data Understanding**

We started our understanding of the data by looking at the data types of the columns. We found that a considerable proportion of those had non-numerical values, such as strings or lists. Similarly, it was important to check the name and description of the features, which allowed us to interpret their meaning and think about potential feature engineering techniques. Careful observation of the data revealed that it required a thorough preprocessing and cleaning in order to be used with ML models.

2. **Data Pre-processing**

For all of the features we explored the number of missing values. If this percentage was relatively low we dropped the feature altogether. In the other cases, we filled the missing values with either the median or the mode, depending on the type of variable. Some of the features had a string format that encoded possible overlapping characteristics, like external or internal insulation. These variables were encoded in a one-hot format, which is very natural for categorical features. We also performed some outlier removal, by clipping the really extreme values of some numerical features. Moreover, we compared the behavior of related features, such as the type of insulation and the thermal conductivity.

3. Modeling Development

We used the models that are known to be more suited for tabular data: xgboost, catboost and lightgbm. The hyperparameters were tuned using manual exploration, as the computational cost of training each model made other approaches like grid search too expensive, and it generally yields a tiny increase in performance. As for the training/validation split, we used a 5-fold cross validation schema.

4. Deployment Strategy

We follow general good practices of software development, encapsulating code in functions and working at different levels of abstraction. The ML procedure was also relevant, preventing data leakage of the validation dataset into the training dataset. We plan to deploy our model using the trending library FastAPI of python, which allows the people in charge of the user interface to easily interact with the model.

CARBON FOOTPRINT LIMITATION

The core of our strategy to limit the carbon footprint was efficient code. At each step we used the best practices with pandas and numpy, avoiding for loops in python and using vectorized operations instead. We also performed a careful feature selection and reduction procedure, keeping only the most relevant features. This helped to keep the computational cost as low as possible and still having a performant model. We also didn't even try using deep learning models, as they are not the best option for tabular data and are very computationally expensive.

CONCLUSION

While at the beginning we focused on trying out basic methods of preprocessing and baseline models, we then decided to perform a detailed feature selection in order to take into account the two relevant factors requested: explainability and carbon footprint. We looked at the feature importance and we chose whether to keep or not each feature based on whether it seemed relevant for the prediction and based on the amount of additional columns that were created. For further improvement, we think it would be interesting to try out other hyperparameters for the models, as we just considered the number of estimators due to training time available. The actual results show a explained variance of 81.27.